Tom Peterka
9700 S. Cass Ave.
Argonne IL 60439
Phone 630-252-7198
Email tpeterka@mcs.anl.gov

Scalable Parallel Building Blocks for Custom Data Analytics

Data analytics consisting of numerical and visual analysis is crucial to the scientific discovery process, but rapidly increasing growth in the size of computational and experimental datasets presents serious challenges for data analytics. The situation is exacerbated by the design constraints of HPC architectures that consistently improve computational rates at the expense of the relative rate of accessing and moving data. The current approach to data analytics, which is based on postprocessing stored data, is untenable in the face of these challenges.

This talk presents an alternative approach to data analytics based on lightweight, custom analysis tasks that are tightly integrated with computational science. Our approach to *in situ* analysis, as this technique is usually called, is based on two conditions. First, the scientist must be willing to take control of her own analysis, ”custom data analytics.” Next, a scalable library of core algorithms for domain decomposition, parallel I/O, and efficient communication is needed, ”scalable parallel building blocks.” Under these conditions, data analytics can truly be executed as any other parallel program, and indeed as a stage of the parallel computation of simulation data or the parallel processing of sensor data. This will bypass large amounts of data movement, reduce the overall turnaround time from hypotheses to conclusions, improve the veracity of those results, and conserve valuable, limited resources.

This vision is supported by an overview of case studies we have performed in this area. Beginning with an summary of parallel visualization algorithms that we have successfully scaled to a large part of leadership architectures such as ANL's Blue Gene/P machine *Intrepid* and ORNL's Cray XT5 machine *Jaguar*:

- Parallel volume rendering
- Parallel image compositing
- Parallel particle tracing

These seemingly unrelated case studies led us to identify common, core capabilities that are needed by these and other parallel analysis tasks. I will discuss these kernels in the context of a prototype library that we are currently developing:

- Domain decomposition
- Parallel I/O
- Scalable communication

The talk concludes with a glimpse at current data analytics techniques that focus more on analysis and less on visualization. Some of these use the above building blocks:

- Information-theoretic Analysis
- Topological Analysis
- Geometric Analysis